山竹台风影响下受灾群众心理状态的台风眼效应
——基于时间与空间维度的微博行为数据分析[1]

朱致琛[1,4] 周意勇[1,4] 王宇宸[1,4] 卢江丰[2,4] 程羽慧[3,4] 何婷婷[2,4] 朱廷劭[1]

（[1] 中国科学院心理研究所行为科学重点实验室，北京 100101）

（[2] 中国科学院心理研究所心理健康重点实验室，北京 100101）

（[3] 中国科学院心理研究所，脑与认知科学国家重点实验室，脑科学与智能技术卓越创新中心，北京 100101）

（[4] 中国科学院大学心理学系，北京 100049）

chinaXiv:201901.00118v1

摘 要 选取山竹台风典型受灾地区(广东)和非受灾地区(安徽)微博用户的行为数据，使用大数据分析的方法从时间和空间两个维度检验心理台风眼效应。结果发现，在时间维度上，受灾地区对台风的关注存在差异，但并没有表现出“高-低-高”的心理台风眼模式，具体来说，受灾地区在台风过境后对其关注程度高于台风过境前，而台风来临前与过境中、来临后与过境中对台风关注程度没有显著差异；在空间维度上，受灾地区和未受灾地区在对台风的关注度上不存在显著差异。本文对研究的局限性进行了分析，以期为未来研究提供相关思考和借鉴。

关键词 山竹台风；时间；空间；微博大数据；心理台风眼效应

分类号 B849

1 前言

山竹台风(Typhoon Mangkhut)作为珠三角地区 37 年来最大的台风灾害(国际编号 1822)，于 2018 年 9 月 16 日在广东泰山海燕镇登陆，短短两天内造成广东、广西、海南、湖南、贵州 5 省(区)，导致近 300 万人受灾，160.1 万人紧急避险转移和安置，1200 余间房屋倒塌，直接经济损失 52 亿元，给国家和人民造成了严重的生命和财产损失(何畅, 2018)。

在造成生命财产损失的同时，自然灾害往往也会给受灾地区人民的心理造成巨大的冲击和伤害。如何正确地认识和描述此次台风给受灾群众带来的心理影响，是解决灾后心理援助问题、后续在其他灾难来临前对人民心态进行调整的基础(沈世林, 张彩云, & 王玉萍, 2014)。本研究致力于次，拟利用

---

[1] 通讯作者：朱廷劭，tszhu@ucas.ac.cn。

chinaXiv:201901.00118v1

大数据分析手段，从时间和空间两个维度出发探究对于台风灾难前后、受灾与非受灾地区人民心理状况的动态变化轨迹。

鉴于台风是可以预测的一种天灾，我们需要了解台风发生前(得知台风消息后到台风登陆前)、台风发生中、台风结束后(台风撤离)的一段时间，相关受灾群众的心理变化过程是怎样的，这可以帮助心理工作人员在不同的时期选择不同的心理干预手段。已经有一些心理学研究，对受灾地区人民心理情况的变化进行了一定的解释。以往研究表明，在灾难发生的时间和空间维度上，存在着一种“心理台风眼效应”(Psychological typhoon eye effect)，即在时间维度上，越接近高风险时段，心理越平静；在空间维度上，越接近高风险地点，心理越平静(李纾, 刘欢, 白新文, 任孝鹏, 郑蕊, 李金珍, 饶俪琳, 汪祚军, 2009)。该理论已经在空间维度(李纾等, 2009)、以及从灾后的时间维度 (时勘, 陈雪峰, 胡卫鹏, 贾建民, 高晶, 李文东, 范红霞, 余俊生, 张丽红, 2003)上得到了证明。但鉴于灾害发生的特点，先前研究从时间维度上难以覆盖到灾难发生前和发生中受灾地区人民或准受灾地区人民的心理状况，而台风这种天灾具有可预测性的特点，可以满足完整地从灾难发生前、中、后探究受灾群众心理变化的需求。另外，台风的可预测性给台风地区的群众提供了一段准备和等待灾难的时期，灾难发生时受灾人民也是有心理预期的，这与地震、洪水等突发性灾害有所区别，可能使得台风受灾群众的心理状态在时间维度产生有别于先前研究的一些特点。

在获取台风受灾群众的心理特征的途径上，“微博”基于其可记录时间地点、数据量大这两个明显的优势，成为本研究很好的数据来源。因此，本研究拟基于台风雷达图，分别从时间和地域两个维度刻画此次山竹台风对受灾地区人民心理特征的影响。具体来说，选取典型受灾区（广东）和非受灾地区（安徽）的人们在台风过境前、中、后三个时间段内的微博行为数据，探究人们对此次台风的关注程度的变化，尝试用心理台风眼效应从时间和空间两个维度来解释人们面对台风这类自然灾害时的心理变化。

本研究以期为帮助心理健康领域的研究者更好地理解可预期的应激事件前后人们的心理状态变化特点，同时在灾前预警、舆情管控和灾后心理救援工作等方面均有一定的实践启示作用。

2 方法

2.1 研究设计

2.1.1 山竹台风影响下人们的关注程度在时间维度上的台风眼效应

目的为分别探究台风发生前中后三个时间段内，受台风影响的广东省人民对山竹台风的关注程度随时间的变化特点。选取时间段为自变量，共有发生前，发生中，发生后三个水平；选取台风山竹的

关注程度(以山竹、台风相关关键词的词频表示)为因变量，对广东省用户的微博行为数据进行分析。研究假设，受灾区在台风发生前、发生中、发生后三个时间段中，人们对山竹台风的关注程度呈现高-低-高变化的特点，表现出时间维度的台风眼效应。

2.1.2　山竹风影响下人们的关注程度在空间维度上的台风眼效应

目的为探究台风发生过程中，受灾地区(广东)与非受灾地区(安徽)对山竹台风的关注程度的差异，因变量与 2.1.1 相同。研究假设，受灾与非受灾地区相比，人们对山竹台风的关注程度存在差异，基于台风眼效应，灾区人民其关注程度反而小于邻近的非灾区人民。

2.2 数据来源

选取此次台风的典型受灾地区和对照地区的微博用户数据。被选取的微博数据需要满足以下条件：

1) 发布该微博的用户，主页-基本资料-个人信息-所在地为：广东或安徽；
2) 微博的发布时间介于台风过境中及其过境前后三天：2018 年 9 月 1 日 0:00 至 2018 年 9 月 30 日 24:00。

本研究爬取的数据内容包括：符合上述(1)所在地要求的用户在上述(2)时间段内所发表的所有微博的文字内容，包括“原创微博”、“转发微博”（转发微博包括转发内容的原文，以及转发者自己写的文字内容）。最终采集到符合条件的微博数据共计 30 万条左右，形式如图 1 所示。

1	ID	用户名	地区	发布时间	微博内容
2	1601518827	卷毛菀	安徽 合肥	1537545600	牛
3	1601518827	卷毛菀	安徽 合肥	1537372800	8:21 6
4	1601518827	卷毛菀	安徽 合肥	1535904000	我在#签到领红包#打卡啦！ 每日签到令
5	1601518827	卷毛菀	安徽 合肥	1536163200	先把耳朵洞钻上。
6	1601518827	卷毛菀	安徽 合肥	1537113600	都是好东西，开一个
7	1601518827	卷毛菀	安徽 合肥	1536595200	超级帅气！这才是男孩子该有的样子
8	1601518827	卷毛菀	安徽 合肥	1537718400	晚一点，早一点都好！不能将就！
9	1601518827	卷毛菀	安徽 合肥	1536249600	两岁多......用的挺好不换......
10	1601518827	卷毛菀	安徽 合肥	1537459200	恶人终有恶报！
11	1601518827	卷毛菀	安徽 合肥	1535904000	不久之后又会有一个女人站出来为自己
12	1601518827	卷毛菀	安徽 合肥	1535817600	转发微博
13	1601518827	卷毛菀	安徽 合肥	1537459200	从来中过奖，月饼节就靠你了。欧哥
14	1601518827	卷毛菀	安徽 合肥	1537286400	大半夜的，想参加个抽奖我容易吗我。
15	1883573041	络文梦剧场	广东 广州	1536163200	只想与你再一起
16	1883573041	络文梦剧场	广东 广州	1538236800	拍了一张照片。　网页链接 #一闪On
17	1883573041	络文梦剧场	广东 广州	1536249600	建设银行白金卡6万8额度什么水平
18	1883573041	络文梦剧场	广东 广州	1536076800	我在#签到领红包#打卡啦！ 每日签到令
19	1883573041	络文梦剧场	广东 广州	1538064000	哈哈哈
20	1883573041	络文梦剧场	广东 广州	1538064000	美丽

图 1. 原始数据样式. notepad

2.3 数据整理

对所采集符合要求的 30 万条左右的数据进行整理，以方便后期分析处理。

1) 将每个省的数据划分为台风过境前、过境中和过境后三个不同的时间段。以台风过境的时间（9 月 16 日—18 日）为中心，将时间戳变量的值转换为 3 个：x < 1537088400 赋为 1，1537088400 <= x

chinaXiv:201901.00118v1

<= 1537174800 赋为 2，1537174800 < x 赋为 3，分别存储为 3 个单独的 xlsx 文件，如“广东-过境前”、“广东-过境中”、“广东-过境后”。

2) 以省为单位建立两个单独的文件夹，以广东省为例，每个文件夹包括“广东省”、“广东-过境前”、“广东-过境中”、“广东-过境后”四个 xlsx 格式的文件，每个文件的首行分别为：用户 ID、用户名、地址、时间戳、微博内容。内容结构如图 2 所示。

1	用户ID	用户名	地址	时间戳	微博内容
2	1694550631	安徽论坛	安徽 合肥	1537286400	台风余威致多地暴雨 凶悍的“山竹”会被除名吗？】17日晚
3	1694550631	安徽论坛	安徽 合肥	1537286400	【炒面打翻在地服务员称重做 不料面中吃出盘子碎片】15
4	2521764812	走拍旅行	安徽 合肥	1538150400	没有领略过坝上的秋色，这旅行的路上多黯然失色。胡守
5	2521764812	走拍旅行	安徽 合肥	1537804800	呈坎晒秋，蜀源古村向日葵，这个秋天，皖南美醉了！星
6	2521764812	走拍旅行	安徽 合肥	1537545600	【发现青海首屈一指的名胜古迹】青海塔尔寺，十万狮子
7	2521764812	走拍旅行	安徽 合肥	1537718400	庐阳中秋夜。金健平/摄 #最美赏月地##走拍旅行#摄影美
8	2521764812	走拍旅行	安徽 合肥	1537459200	#走拍旅行#中国最好玩的地方开封小宋城，好玩好吃的地
9	2521764812	走拍旅行	安徽 合肥	1537372800	彼岸花中国花语:优美纯洁朝鲜花语:相互思念日本花语:悲
10	2521764812	走拍旅行	安徽 合肥	1538236800	一个人，要在内心里给最喜欢的人，给喜欢的地方，喜欢
11	2521764812	走拍旅行	安徽 合肥	1537286400	【发现沉睡的百年古镇，比凤凰古镇还淳朴】坐落于福建
12	2521764812	走拍旅行	安徽 合肥	1537372800	找一个喜欢的人，一起去看川西秋色吧！徐旭/摄 #走拍旅
13	2521764812	走拍旅行	安徽 合肥	1537977600	中国最美滩涂，霞浦。拍摄地：下青山落日，江村s弯，杨
14	2521764812	走拍旅行	安徽 合肥	1537718400	今夜月明人尽望，不知秋思落谁家？图/徐国友 #最美赏月
15	2521764812	走拍旅行	安徽 合肥	1537286400	台风后，今天下午四点，安徽合肥东方的天空出现巨大的“
16	1773263881	qa1ws23	安徽 芜湖	1537632000	//@午后狂睡: 英雄。//@天凶: //@庆丰:正能量
17	1773263881	qa1ws23	安徽 芜湖	1537632000	//@阴楼孤魂: 转发微博
18	1773263881	qa1ws23	安徽 芜湖	1538064000	//@老白喵喵喵: //@红麟or阿莹-废材组 ://@Mia_薛定谔的
19	1773263881	qa1ws23	安徽 芜湖	1538150400	//@阴楼孤魂: //@孔令旗的地盘://@看看底牌: 触目惊心//
20	1773263881	qa1ws23	安徽 芜湖	1537459200	//@那谁家的兔纸: //@Lilys :又蠢又坏//@只谈风月猫 ://@
21	1773263881	qa1ws23	安徽 芜湖	1537286400	//@美食家大雄: 每天送孩子上学放学都是早晚高峰，司机
22	1773263881	qa1ws23	安徽 芜湖	1537891200	//@饼干姐: //@暴雨雷霆: 我以后讲课都要提醒民警，梨视
23	1773263881	qa1ws23	安徽 芜湖	1537977600	//@杭之冯玥均建国后成精: 所以还是要背锅啊。//@来去
24	1773263881	qa1ws23	安徽 芜湖	1538150400	//@阴楼孤魂: //@宽容公正麦卡锡: @押沙龙??//@華朱咖
25	1773263881	qa1ws23	安徽 芜湖	1537804800	//@泉涸-鱼相与处于陆: 恐怖片d(?д??)
26	1773263881	qa1ws23	安徽 芜湖	1537372800	//@天闻角川: DO IT.//@重工组长于彦舒: 哈哈哈哈哈//@
27	1773263881	qa1ws23	安徽 芜湖	1537977600	//@孔令旗的地盘: 现在社会给家长、老师教育孩子的权限

安徽-过境后

图 2. “安徽-过境后”.xlsx 内容结构

2.4 数据分析

基于台风词汇库，选取 29 个与台风相关的词汇，对每条“微博内容”的词频数进行统计。

1) 将“用户 ID”和“时间戳”均相同且“地址”为“广东”的“微博内容”的台风相关词频数相加得到若干个频数 cnti，每个 cnti 只对应一个“用户 ID”，从而只对应一个“地址”，每个“用户 ID”不一定只对应一个 cnti；计算“用户 ID”和“时间戳”均相同的“微博内容”的总汉字数 ni。用 cnti 除以相应的 ni，得到该用户在该时间段微博内容中台风相关词语的频率 pi。将“时间戳”为 1 的 pi 放入列表 list1；将“时间戳”为 2 的 pi 放入列表 list2；将“时间戳”为 3 的 pi 放入列表 list3。

2) 将“用户 ID”相同的“微博内容”的台风相关词频数相加得到若干个频数 cnti，每个 cnti 只对应一个 ID，每个 ID 也只对应一个 cnti；计算“用户 ID”相同的“微博内容”的总字数 ni。用 cnti 除以相应的 ni，得到该用户在整个台风时间段微博内容中台风相关词语的频率 pi。将“地址”为广东的 pi 放入 list1；将 SPACE 为安徽的 pi 放入 list2。

3) 对第一步和第二步的 list，分别做两两差异检验，结果示意图如图 3。

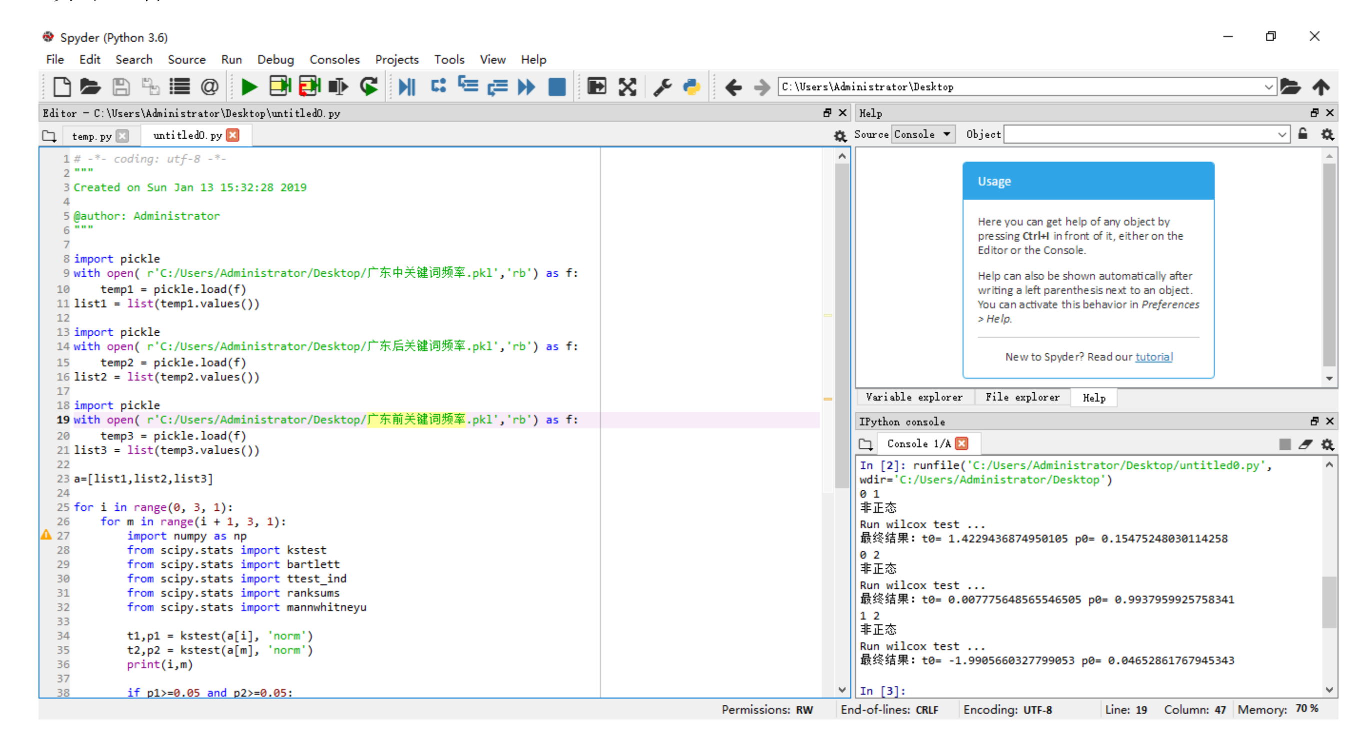
chinaXiv:201901.00118v1

3 结果

本研究利用大数据分析手段，从时间和空间两个维度出发探究了台风灾难发生前后、受灾与非受灾地区人民心理状况的动态变化轨迹。

在时间维度上，广东省用户微博行为数据表明，台风来临前和台风来临后的“台风相关词汇”频率存在显著差异($t=-1.99$; $p=0.05$)，台风来临前的频率显著低于台风来临后，但并没有表现出“高-低-高”的心理台风眼模式。我们认为有两个可能的原因：一方面，广东地区由于地理位置的特殊性，台风经验和台风隐患较多，故而在发布台风预警时人们对台风的关注并不会显著增加；另一方面，随着对台风的认识和应急措施的不断完善，在台风登陆后，受灾区人们的台风体验和新闻报道等的不断更进，进而人们对其的关注会有所增加。

在空间维度上，广东省和安徽省的微博行为数据表明，受灾地区和非受灾地区的对台风的关注程度在台风来临前中后整个时段内，都没有显著差异。这可能是因为在微博上，台风的相关推送和报道覆盖性很强，这导致非受灾地区的微博用户也会出现及时更进台风的相关报道，在关注度上差异不显著。



chinaXiv:201901.00118v1

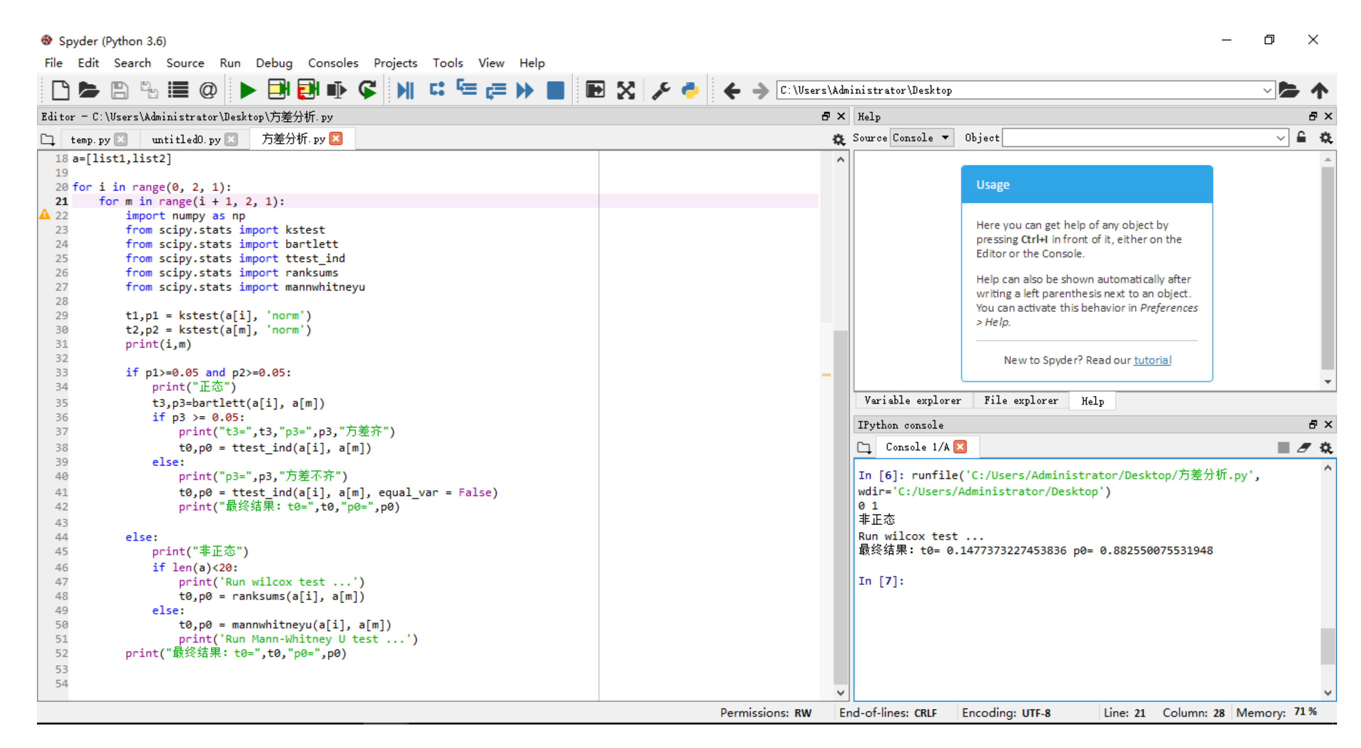


图 3. 方差分析结果示意图

4 讨论

本研究通过微博行为数据分析的方法，探究受灾地区及其邻近地区“山竹台风”的关注度是否在时间和空间上存在台风眼效应，结果并未发现时间和空间上的台风眼效应。除了上述可能的原因外，也可能是研究方法的缺陷，由于本研究反应关注度的方法是计算提及关键词的词频，可能“台风相关词汇”的词频并不能很好地反映对某件事物的关注程度，例如用户可能表达了台风相关内容但并未提及相关词汇。在后续研究中，应该考虑使用更丰富的指标反应关注程度。

本研究具有重要的理论和实际意义。在理论上，能够帮助心理健康领域的研究者更好地理解可预期的应激事件前后(如台风、地震、泥石流等自然灾害)，受影响人群的心理状态变化特点。同时在实践上，本研究对灾前预警、舆情管控和灾后心理救援工作等方面均有一定的启示作用。

同时，本研究也存在一些局限性。首先上文已经提及，可能“台风相关词汇”词频并不能很好地反映对某件事物的关注程度，后续研究中，应该考虑使用更丰富的指标反应关注程度，例如“负性情绪相关词汇”的词频。其次，本研究没有对受灾经验维度进行控制。本次山竹台风受到波及的地区中也存在着一些有趣的差异点，比如有些省份从前并没有台风(例如湖南)，而有些省份则曾经受到，甚至多次受到台风的打击(例如广东)。是否有过受灾经验，对相关地区的人民面对灾害时的心理特征可能会产生一定的影响 (健哉, 2008; 雷晓敏, 2011)。未来研究或可致力于此，剥离开受灾经验的潜在影响，更好地揭示受灾地区人们的心理活动状态。另外，本研究并未囊括菲律宾这一受灾重区。本次山竹台

chinaXiv:201901.00118v1

风在国际换日线以东海域形成后，不断西进，先后影响到了毗邻的中国和菲律宾地区，且菲律宾地区遇难人数远超于中国。本研究对于数据的选取囿于微博，而不是更具国际通识性的数据源，如气象局资料或新闻报道等，这使得研究样本局限于中国地区。对比此次台风对菲律宾与中国地区造成影响的差异性，对灾前预警和舆情管控等方面有着重要的启示，比如台风率先登陆广东，可能对毗邻的菲律宾具有警示作用，其民众对台风的关注度和焦虑程度会呈现出“提前”的效果，而不是典型的“台风眼”效应。未来研究或可尝试对样本量进行扩大，继续深入分析台风类自然灾害事件对人们心理状态的影响。

参考文献

何畅. (2018). 强台风“山竹”带给我们的反思. *深圳商报.*

雷晓敏. (2011). 中国民众抗震自救能力研究——以日本国民消解震灾经验为例. *战略决策研究, 02*(3), 51-55.

李纾, 刘欢, 白新文, 任孝鹏, 郑蕊, & 李金珍等. (2009). 汶川“5.12”地震中的“心理台风眼”效应. *科技导报, 27*(3), 87-98.

沈世林, 张彩云, & 王玉萍. (2014). 重大自然灾害创伤后心理应激障碍研究现状. *卫生职业教育, 32*(12), 158-159.

时勘, 陈雪峰, 胡卫鹏, 贾建民, 高晶, 李文东, 范红霞, 余俊生, 张丽红. (2003). 北京市民对 sars 疫情的风险认知特征. *人口研究, 27*(4), 44-48.

健哉. (2008). 日本的抗震救灾经验. *中国减灾, 2008*(6), 50-51.

chinaXiv:201901.00118v1

The Effect of Typhoon Eye on the Psychological State of the Victims under the Impact of Typhoon Mangosteen: Analysis of Microblog Behavioral Data Based on Time and Spatial Dimensions

Zhu Zhichen[1, 4] Zhou Yiyong[1, 4] Wang Yuchen[1, 4] Lu Jiangfeng[2, 4] Cheng Yuhui[3, 4] He Tingting[2, 4] Zhu Tingshao[1]

([1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China)

([2] CAS Key Laboratory of Mental Health, Institute of Psychology, Beijing 100101, China)

([3] State Key Laboratory of Brain and Cognitive Science, CAS Center for Excellence in Brain Science and

Intelligence Technology, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China)
([4] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract

The Microblog behavioral data of typical disaster-stricken areas under the Impact of Typhoon Mangosteen (Guangdong) and non-disaster areas (Anhui) were selected to test the psychological typhoon eye effect from two dimensions of time and space with the method of big data analysis. The results show that there are differences of attention degree in the time dimension, but there is no "high-low-high" psychological typhoon eye pattern in the affected areas. Specifically, the attention of the affected areas after the transit of typhoon is higher than that before the transit of typhoon, while there is no significant difference of the attention of typhoon between the time periods before, during and after the transit. There is no significant difference between the disaster-stricken areas and the non-disaster areas in the dimension of spatial. Limitations of the study are mainly analyzed in order to provide relevant thinking and reference for future research.

Key words: Typhoon Mangosteen; Time; Spatial; Big Data on Microblog; Psychological Eye Effect of Typhoon

chinaXiv:201901.00118v1

附录 1 台风词库（22 个）

强台风，台风，山竹，应急预案，暴雨，超强台风，气象专家，强风，灾害，防护，危害，降雨，检测，预报，天气，台风眼，风力等级，平均风速，气旋，灾害性天气，气压，路径，亮温，风力，风速，气压梯度，天气图，天气报告，高压

chinaXiv:201901.00118v1

附录 2 源代码

（1） 计算台风词频代码

```
# -*- coding: utf-8 -*-
##################对数据的时间进行标签化
import openpyxl
from collections import defaultdict
wb = openpyxl.load_workbook(r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop\安徽-过境后.xlsx')
wss = [wb[i] for i in wb.sheetnames]

fail_datas = []
for ws in wss:
    fail_data = []
    try:
        for i in range(2,ws.max_row+1):
            if not ws.cell(row = i,column = 4).value:
                fail_data.append(str(wb.sheetnames[wss.index(ws)])+str(i))
            elif int(ws.cell(row = i,column = 4).value) < 1537088400:
                ws.cell(row = i,column = 4).value = 1
            elif int(ws.cell(row = i,column = 4).value) >= 1537088400 and int(ws.cell(row = i,column = 4).value)<=1537174800:
                ws.cell(row = i,column = 4).value = 2
            else:
                ws.cell(row = i,column = 4).value = 3
    except Exception as e:
        print(e)
        fail_data.append(str(wb.sheetnames[wss.index(ws)])+str(i))
    fail_datas.append(fail_data)

wb.save(r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop\安徽-过境后_带标签.xlsx')

filename_list = ['广东-过境前.xlsx','广东-过境中.xlsx','广东-过境后.xlsx','安徽-过境前.xlsx','安徽-过境中.xlsx','安徽-过境后.xlsx']
######创建一个 key 为 id，值为 content 的字典。用于读取所有的文件中的评论
def id_content(filename):
    wb = openpyxl.load_workbook(filename)
    wss = [wb[i] for i in wb.sheetnames]
    i = 1
    ws = wss[0]
    id_content_dict = defaultdict(list)
    #ID = []
    #for j in range(1,ws.max_row+1):
    #    ID.append(ws.cell(row=j,column=1))
```


chinaXiv:201901.00118v1

```
    try:
        while(i<=ws.max_row):
            content_each = [ws.cell(row=i,column=5).value]
            m = i+1
    #           for m in range(i+1,i+600):
    #               if ws.cell(row=i,column=1).value == ws.cell(row=m,column=1).value:
            while ws.cell(row=i,column=1).value == ws.cell(row=m,column=1).value:
                content_each.append(ws.cell(row=m,column=5).value)
                m = m+1
            id_content_dict[ws.cell(row=i,column=1).value] = content_each
    #           print(content_each)
            i = m+1
    #           print(i)
    except Exception as e:
        print(e)
    return id_content_dict
#a=id_content(r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop\广东-过境前.xlsx')

import os
import pickle
path = r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop'
for i in filename_list:
    file_path = os.path.join(path,i)
    out_file_path = os.path.join(path,i.split('.')[0]+'.pkl')
    id_content_dic = id_content(file_path)
    with open(out_file_path,'wb') as f:
        pickle.dump(id_content_dic,f)
#####将每个省的前中后时期评论内容变为 id_content 的词典

####拼接同一时间的两个省的评论的内容
def uni_content(file_name):
    path = r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop'
    two_all_id_content_dict = defaultdict(dict)
    sec_file_name = filename_list[filename_list.index(file_name)+3]
    pkl_filename_1 = os.path.join(path,i.split('.')[0]+'.pkl')
    pkl_filename_2 = os.path.join(path,sec_file_name.split('.')[0]+'.pkl')
    with open(pkl_filename_1,'rb') as f1:
        id_content_dict_1 = pickle.load(f1)
    with open(pkl_filename_2,'rb') as f2:
        id_content_dict_2 = pickle.load(f2)
    for j in list(id_content_dict_1.keys()):
        two_all_id_content_dict[j] = id_content_dict_1[j]
    for j in list(id_content_dict_1.keys()):
        for m in list(id_content_dict_2.keys()):
```


chinaXiv:201901.00118v1

```
            if j == m:
                uni_content = id_content_dict_1[j]+id_content_dict_2[m]
                two_all_id_content_dict[j] = uni_content
            else:
                two_all_id_content_dict[m] = id_content_dict_2[m]
    return two_all_id_content_dict,sec_file_name

####拼接同一个省三个时间段的评论的内容
def uni_content_three(file_name):
    path = r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop'
    three_content_list = []
    sec_file_name = filename_list[filename_list.index(file_name)+1]
    third_file_name = filename_list[filename_list.index(file_name)+2]
    pkl_filename_1 = os.path.join(path,i.split('.')[0]+'.pkl')
    pkl_filename_2 = os.path.join(path,sec_file_name.split('.')[0]+'.pkl')
    pkl_filename_3 = os.path.join(path,third_file_name.split('.')[0]+'.pkl')
    with open(pkl_filename_1,'rb') as f1:
        id_content_dict_1 = pickle.load(f1)
    with open(pkl_filename_2,'rb') as f2:
        id_content_dict_2 = pickle.load(f2)
    with open(pkl_filename_3,'rb') as f3:
        id_content_dict_3 = pickle.load(f3)
    content_1 = list(id_content_dict_1.values())
    content_2 = list(id_content_dict_2.values())
    content_3 = list(id_content_dict_3.values())
    three_content_list = content_1+content_2+content_3
    return three_content_list

def keyword_fre(keywords,content_dict):####计算出两个省各个时间段的关键词占总数的比例
    keywords_fre_dict = {}.fromkeys(keywords,0)
    content_to_str = [str(j) for i in list(content_dict.values()) for j in i ]
    content_list = [''.join(i) for i in content_to_str ]
    content_str =''.join(content_list)
    for i in keywords:
        fre_time = content_str.count(i)####关键词出现的次数
        fre = fre_time *len(i)/len(content_str)
        keywords_fre_dict[i] = fre
    return keywords_fre_dict

def keyword_fre_onecity(keywords,content_list):####计算出一个省所有时间段的关键词占总数的比例
    keywords_fre_dict = {}.fromkeys(keywords,0)
    content_to_str = [str(j) for i in content_list for j in i ]
    content_list_1 = [''.join(i) for i in content_to_str ]
    content_str =''.join(content_list_1)
```


chinaXiv:201901.00118v1

```
    for i in keywords:
        fre_time = content_str.count(i)####关键词出现的次数
        fre = fre_time *len(i)/len(content_str)
        keywords_fre_dict[i] = fre
    return keywords_fre_dict

def main_2():
    str1 = '强台风，台风，山竹，应急预案，暴雨，超强台风，气象专家，强风，灾害，防护，危害，降雨， 检测，预报，天气，台风眼，风力等级，平均风速，气旋，灾害性天气，气压，路径，亮温，风力，风速，气压梯度，天气图，天气报告，高压'
    str2 = str1.split("，")
    key_word_taifeng = [i.strip() for i in str2]
    path = r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop'
    for i in filename_list[0:6:3]:
        all_content = uni_content_three(i)
        keyword_frequency = keyword_fre_onecity(key_word_taifeng,all_content)
        file_name = i.split('-')[0]+'关键词频率.pkl'
        out_file_path = os.path.join(path,file_name)
        with open(out_file_path,'wb') as f:
            pickle.dump(keyword_frequency,f)
main_2()

def main_1():
    str1 = '强台风，台风，山竹，应急预案，暴雨，超强台风，气象专家，强风，灾害，防护，危害，降雨， 检测，预报，天气，台风眼，风力等级，平均风速，气旋，灾害性天气，气压，路径，亮温，风力，风速，气压梯度，天气图，天气报告，高压'
    str2 = str1.split("，")
    key_word_taifeng = [i.strip() for i in str2]
    path = r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop'
    for i in filename_list[0:3]:
        two_all_content,second_file_name = uni_content(i)
        keyword_frequency = keyword_fre(key_word_taifeng,two_all_content)
        file_name = i.split('-')[0]+second_file_name.split('-')[0]+i.split('.')[0][-1]+'关键词频率.pkl'
        out_file_path = os.path.join(path,file_name)
        with open(out_file_path,'wb') as f:
            pickle.dump(keyword_frequency,f)
main_1()

test_dic,test_filename = uni_content('广东-过境中.xlsx')
with open( r'C:\Users\lbj.WIN-40GLMDFSSV7\Desktop\安徽关键词频率.pkl','rb') as f:#######读取存的关键词频率字典
    AH = pickle.load(f)
```

chinaXiv:201901.00118v1

（2） 方差分析代码

```
import pickle
import numpy as np
from scipy.stats import kstest
from scipy.stats import bartlett
from scipy.stats import ttest_ind
from scipy.stats import ranksums
from scipy.stats import mannwhitneyu
###将文件转化为 list
with open( r'C:/Users/Administrator/Desktop/广东中关键词频率.pkl','rb') as f:
    temp1 = pickle.load(f)
list1 = list(temp1.values())
with open( r'C:/Users/Administrator/Desktop/广东后关键词频率.pkl','rb') as f:
    temp2 = pickle.load(f)
list2 = list(temp2.values())
with open( r'C:/Users/Administrator/Desktop/广东前关键词频率.pkl','rb') as f:
    temp3 = pickle.load(f)
list3 = list(temp3.values())
a=[list1,list2,list3]
###list 两两之间进行差异检验
for i in range(0, 3, 1):
    for m in range(i + 1, 3, 1):
        t1,p1 = kstest(a[i], 'norm')
        t2,p2 = kstest(a[m], 'norm')
        print(i,m)
        if p1>=0.05 and p2>=0.05:
            print("正态")
            t3,p3=bartlett(a[i], a[m])
            if p3 >= 0.05:
                print("t3=",t3,"p3=",p3,"方差齐")
                t0,p0 = ttest_ind(a[i], a[m])
            else:
                print("p3=",p3,"方差不齐")
                t0,p0 = ttest_ind(a[i], a[m], equal_var = False)
                print("最终结果：t0=",t0,"p0=",p0)
        else:
            print("非正态")
            if len(a)<20:
                print('Run wilcox test ...')
                t0,p0 = ranksums(a[i], a[m])
            else:
                t0,p0 = mannwhitneyu(a[i], a[m])
                print('Run Mann-Whitney U test ...')
        print("最终结果：t0=",t0,"p0=",p0)
```


chinaXiv:201901.00118v1